Volume 9, Number 1 January 1996

JTSTEB 9(1) 1-158 (1996)

ISSN 0894-9867

Journal of Traumatic Stress

PLENUM PRESS • NEW YORK-LONDON

Journal of Traumatic Stress, Vol. 9, No. 1, 1996

The Utility of the SCL-90-R for the Diagnosis of War-Zone Related Posttraumatic Stress Disorder

Frank W. Weathers,[1] Brett T. Litz,[1] Terence M. Keane,[1] Debra S. Herman,[1] Howard R. Steinberg,[1] Jennifer A. Huska,[1] and Helena C. Kraemer[2]

*A scale for assessing war-zone-related posttraumatic stress disorder (WZ-PTSD scale) was derived from the Symptom Checklist-90-R by identifying items that best discriminated Vietnam theater veterans with and without PTSD (*N *= 202). The 25-item WZ-PTSD scale had excellent internal consistency, and signal detection analyses revealed that its diagnostic utility was comparable to or exceeded that of several established PTSD scales and measures of global distress. In a cross-validation sample (*N *= 99), the diagnostic utility of the WZ-PTSD scale was stable, whereas other PTSD scales performed more poorly. The WZ-PTSD scale appears to be a valuable new measure of PTSD that can be particularly useful in archival data sets or in any situation where other PTSD measures are not available.*

KEY WORDS: PTSD; assessment; psychometric; signal detection; SCL-90-R.

A number of reliable and valid questionnaires are now available for assessing posttraumatic stress disorder (PTSD), including the Impact of Event Scale (IES; Horowitz, Wilner, & Alvarez, 1979; Zilberg, Weiss, & Horowitz, 1982), the PK scale of the MMPI and the MMPI-2 (Keane, Malloy, & Fairbank, 1984; see Litz et al., 1991, and Lyons & Keane, 1992), the Mississippi Scale for Combat-Related PTSD (Mississippi Scale; Keane, Caddell, & Taylor, 1988), the Penn Inventory (Hammarberg, 1992), the

[1] Boston Department of Veterans Affairs Medical Center, Boston, Massachusetts 02130; and Tufts University School of Medicine, Medford, Massachusetts 02155.
[2] Stanford University School of Medicine, Stanford, California 94305.

PTSD Symptom Scale (PSS; Foa, Riggs, Dancu, & Rothbaum, 1993), and the PTSD Checklist (PCL; Weathers, Litz, Herman, Huska, & Keane, 1993).

These scales are easy to administer and score, they provide a continuous measure of the severity of PTSD symptoms, and they can be used to predict diagnostic status once optimal cutoff scores have been established. For these reasons, information from such questionnaires has become an integral component of a multimethod assessment of PTSD (see Keane, Wolfe, & Taylor, 1987; Kulka et al., 1990). However, despite considerable progress in the psychometric assessment of PTSD, these scales have not been widely adopted outside of clinical research settings. This is unfortunate because recent research has documented the high prevalence of psychological trauma in various treatment-seeking populations (e.g., Blake, Keane et al., 1990; Davidson & Smith, 1990), suggesting the need for more widespread screening for PTSD in settings where trauma-related problems may otherwise go undetected and untreated.

In developing new self-report measures of PTSD, investigators have taken two different approaches to scale construction: a rational approach in which item content reflects the clinical phenomenology of PTSD, and an empirical approach in which items are chosen from an existing non-PTSD scale solely on the basis of their statistical ability to discriminate PTSD cases from non-cases. A significant advantage of rationally derived scales, such as the IES, the Mississippi Scale, and the Penn Inventory, is that their close correspondence with the core symptoms and associated features of PTSD enhances validity. However, it is not always possible to use these scales. For example, since these measures are so new, they are not available for secondary analyses in archival data sets. Also, in some clinical or research settings, existing assessment batteries are already so extensive that adding PTSD questionnaires might place an undue burden on examinees or clinicians.

In situations where rational scales cannot be used, empirically derived scales can be a viable alternative, thereby significantly increasing access to PTSD measures. For example, the PK scale can be obtained from any assessment battery containing the MMPI or MMPI-2, making it available both in archival data sets collected before specialized PTSD measures were available, as well as in clinical or research settings where rationally derived PTSD scales are not administered. A recently completed study by Spiro, Schnurr, and Aldwin (1994) illustrates the use of the PK Scale to measure combat-related PTSD in an archival data set. A second advantage of using empirically derived scales is that parent instruments such as the MMPI can provide additional valuable information regarding such issues as comorbid problems or response validity.

Another instrument besides the MMPI that might serve well as a source for an empirically derived PTSD scale is the Symptom Checklist-90-Revised (SCL-90-R; Derogatis, 1983). The SCL-90-R has several advantages in this regard: (a) it is one of the most widely used general measures of psychopathology; (b) it assesses a wide range of psychopathology, so it can provide information on comorbid problems such as depression, anxiety, and hostility; (c) it is relatively brief to administer, often taking only 15-20 min, and it is available in a 53-item version, known as the Brief Symptom Inventory (BSI), which requires even less time to administer; and (d) it is available in many archival databases, making it possible to measure PTSD in a wide range of clinical and research settings.

Recently, Saunders, Arata, & Kilpatrick (1990) used the empirical approach to develop a measure of crime-related PTSD based on the SCL-90-R. They compared female crime victims with and without PTSD on each of the 90 items on the SCL-90-R, identifying 28 items that discriminated the two groups beyond the .0001 level. These 28 items appeared to form a unidimensional scale, as indicated by a very high degree of internal consistency (Cronbach's alpha = .93). Also, regarding diagnostic utility, when scale scores were used in a discriminant function analysis to predict diagnostic status on the Diagnostic Interview Schedule (DIS; Robins, Helzer, Croughan, & Ratcliff, 1981), 89% of the subjects were classified correctly, indicating strong diagnostic utility.

In the present study, we sought to evaluate the efficacy of the Saunders et al. (1990) PTSD scale for assessing war-zone-related PTSD. (Throughout this paper we use the term "war-zone-related PTSD" instead of the more restrictive "combat-related PTSD" in recognition of the fact that many war-zone stressors not related directly to combat can elicit PTSD symptoms.) Although we expected similarities in the patterns of PTSD symptoms endorsed by crime victims and combat veterans, we anticipated the possibility that these two populations might be sufficiently different to warrant the development of a separate scale for war-zone-related PTSD. Therefore, we began by examining individual SCL-90-R items to determine if the items constituting the Saunders et al. (1990) scale were the same items that best discriminated combat veterans with and without PTSD.

A second purpose for the present study was to compare the diagnostic utility of several PTSD scales, using signal detection methodology outlined by Kraemer (1992). Although some individual PTSD scales have been the focus of extensive research (see Watson, 1990, for a review), few studies have evaluated the relative diagnostic utility of these scales. Such a comparison would be valuable in guiding clinicians and investigators in their selection of optimal tests for various clinical and research purposes (e.g. screening, differential diagnosis).

Measures of the diagnostic utility of a questionnaire are calculated from 2 × 2 contingency tables created by: (a) selecting a cutoff score on a questionnaire and dichotomizing a sample of subjects into test-positives (subjects at or above the cutoff score) and test-negatives (subjects below the cutoff score); and (b) comparing status on the test with diagnostic status based on a well-accepted diagnostic procedure (or "gold standard"), in this case a structured interview. Questionnaires typically yield a range of possible cutoff scores, and the diagnostic utility of each cutoff can be evaluated separately. In the present study we evaluated the diagnostic utility of each possible cutoff score on each of the questionnaires considered.

In discussing diagnostic utility, Kraemer (1992) distinguished between measures of test performance and measures of test quality. Commonly reported measures of test performance include sensitivity (the probability that subjects with a positive diagnosis receive a positive test), specificity (the probability that subjects with a negative diagnosis receive a negative test), and efficiency (probability that the test and the diagnosis agree). Evaluating the performance of each cutoff score on the same questionnaire creates what Kraemer (1992) calls a "nested" family of tests. Within this family of tests, there is a tradeoff between sensitivity and specificity as increasingly lenient or stringent cutoff scores are considered. Lenient cutoffs have higher sensitivity but lower specificity relative to stringent cutoffs. Highly efficient cutoff scores tend to have a balance between sensitivity and specificity. The performance of a questionnaire across the entire range of cutoff scores can be depicted graphically by plotting sensitivity against specificity. Such a graph results in a Receiver Operating Characteristic (ROC) curve (cf. Hanley & McNeil, 1982; Swets & Pickett, 1982).

Although sensitivity, specificity, and efficiency depict relationships between test and diagnosis, these measures of test performance are uncalibrated and thus are ambiguous indicators of diagnostic utility unless adjusted for chance agreement between diagnosis and test (Kraemer,1992). Utilizing weighted kappa coefficients, Kraemer (1992) proposes indices reflecting the quality of sensitivity [$\kappa(1)$], specificity[$\kappa(0)$], and efficiency[$\kappa(.5)$]. These quality indices are calibrated measures in that they have fixed endpoints, with a value of .00 equivalent to chance agreement between the diagnosis and the test, and a value of 1.00 equivalent to perfect agreement.

By comparing the quality indices for different cutoff scores on the same questionnaire, investigators can easily identify the optimally sensitive, specific, and efficient cutoffs. These typically correspond to different scores and are useful for different assessment needs: Sensitive cutoffs are more lenient, increasing true positives and reducing false negatives; specific cutoffs are more stringent, increasing true negatives and reducing false positives; and efficient cutoffs optimize the number of agreements between the diagnosis and the test.

In the present research, we focused on identifying optimally efficient cutoffs because we were concerned primarily with differential diagnosis, or the extent to which various questionnaires could accurately predict a PTSD diagnosis. However, we also determined the optimally sensitive and specific cutoffs for each questionnaire. Optimally sensitive cutoffs are useful for screening, and optimally specific cutoffs are useful for making definitive diagnoses (Kraemer, 1992).

This report describes the results of two studies, a derivation study and a cross-validation study. For item analyses in the derivation study (Study 1), we divided a sample of Vietnam theater veterans into two matched subsamples. Within each subsample, we identified SCL-90-R items that best discriminated veterans with and without PTSD. Finding only partial overlap between these items and the items constituting the Saunders et al. (1990) PTSD scale, we created a 25-item War-Zone-Related PTSD scale (WZ-PTSD). Next, we compared the diagnostic utility of the WZ-PTSD scale with that of other questionnaire measures of PTSD and global measures of distress, using the entire derivation sample. In the cross-validation study (Study 2), we compared the WZ-PTSD scale with the other measures in an independent sample of Vietnam theater veterans.

Study 1

Method

Subjects

Subjects in the derivation sample were 202 male Vietnam theater veterans who had contacted the National Center for PTSD from October, 1989 to October, 1991 either to obtain clinical services (63%) or to participate in research (37%). Subjects were primarily White (85%) and African-American (11%), had at least a high school education (91%), and were typically veterans of the Army (50%) and Marines (38%). Mean age for the sample was 43.5 (*SD* = 2.94). Subjects were classified either as PTSD (67%) or non-PTSD (33%) on the basis of a structured diagnostic interview. PTSD and non-PTSD subjects did not differ significantly on any of the demographic measures.

For the purpose of item analyses, two comparable subsamples of 101 subjects were created by: (a) matching subjects on diagnosis (PTSD/non-PTSD), source of diagnosis (CAPS/SCID; see Measures), reason for contact (clinical/research), and PTSD symptom severity (Mississippi Scale scores; see Measures); then (b) randomly assigning members of each matched pair to different subsamples. The two subsamples did not differ significantly on any of the matching variables or on any of the demographic measures.

Measures

Diagnostic interviews. To determine PTSD diagnoses, subjects were administered either the PTSD module of the Structured Clinical Interview for DSM-III-R (Spitzer, Williams, Gibbon, & First, 1990) or the Clinician-Administered PTSD Scale (CAPS; Blake, Weathers et al., 1990). The SCID has been the most widely used interview for diagnosing PTSD, and several studies have documented its reliability and validity (e.g., Keane, Kolb, & Thomas, 1990; Kulka et al., 1990; McFall, Smith, Roszell, Tarver, & Malas, 1990). The CAPS is a new structured interview for PTSD that has excellent psychometric properties, as described in a preliminary report (Blake, Weathers et al., 1990) and in a recently completed, large-scale investigation (Weathers et al., 1992; Weathers & Litz, 1994).

A diagnosis of PTSD was made according to DSM-III-R criteria, using information about the presence or absence of symptoms derived from the SCID or the CAPS. The stressor criterion was exposure to war-zone stress as measured by the Combat Exposure Scale (Keane et al., 1989) and by an open-ended interview regarding service in the Vietnam theater. Of the 202 subjects in the derivation sample, 151 (75%) were diagnosed using the SCID PTSD module, and 51 (25%) were diagnosed using the CAPS.

Questionnaires. All 202 subjects completed the SCL-90-R (Derogatis, 1983) and the Mississippi Scale (Keane et al., 1988), and 166 subjects completed either the MMPI (Hathaway & McKinley, 1983) or the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). The Mississippi Scale is a 35-item measure of PTSD and associated symptoms. It is the most widely used questionnaire for assessing combat-related PTSD, and in the National Vietnam Veterans Readjustment Study (NVVRS; Kulka et al., 1990) it emerged as the best predictor of a PTSD diagnosis.

The PK scale (Keane et al., 1984), which can be obtained from either the MMPI or the MMPI-2 or administered as a stand-alone scale (Herman, Weathers, Litz, Joaquim, & Keane, 1993), also has been utilized extensively in clinical investigations of combat-related PTSD. The original PK scale, derived from the MMPI, consisted of 49 items, including 3 duplicate items. In the MMPI-2, the duplicate items were eliminated, creating a 46-item PK scale (see Lyons & Keane, 1992). In the present study, we used only the 46 nonduplicated items for all subjects, regardless of whether they completed the MMPI (n = 59), the MMPI-2 (n = 107), or the stand-alone version of the PK scale (n = 32).

Item Analyses

SCL-90-R items that best discriminated combat veterans with and without PTSD were identified by means of t-tests conducted on all 90 items. Separate analyses comparing PTSD and non-PTSD subjects were conducted within each of the two subsamples. Only 11 of the 28 items on the CR-PTSD scale were among the 30 items that best discriminated PTSD and non-PTSD subjects in both subsamples. Given this modest number of overlapping items, we decided to construct a new scale to assess war-zone related PTSD and to compare its diagnostic utility with that of the CR-PTSD scale.

The item analyses revealed that PTSD and non-PTSD subjects differed significantly on most of the SCL-90-R items (82 items in one subsample and 85 items in the other). Therefore, in order to develop a scale with maximal power of discrimination, we struck a compromise between retaining only those items that most robustly discriminated veterans with and without PTSD and retaining a sufficient number of items to form a reliable scale. We decided to retain items for the WZ-PTSD scale only if the eta^2 for the difference between PTSD and non-PTSD subjects was .15 or greater in both subsamples, a value which with this sample size corresponds to $p < .0001$. [The eta^2 statistic is a measure of the proportion of variance accounted for by an independent variable, in this case diagnostic group (PTSD vs. non-PTSD; see Hays 1988).]

A total of 25 items met or exceeded the selection criterion, and these items constitute the WZ-PTSD scale. (SCL-90-R items included in the WZ-PTSD scale are 2* 3, 23*, 24*, 29*, 30*, 32* 39, 43*, 44*, 50* 55*, 57*, 59*, 66, 67*, 70*, 71, 72*, 77*, 78*, 79*, 86, 89*, and 90*. Items marked with an asterisk also appear on the BSI.) Since the items on the WZ-PTSD scale showed equally robust discrimination in both subsamples, subsequent analyses were based on the entire derivation sample, collapsed across subsamples. The internal consistency of the WZ-PTSD scale was quite high, suggesting that it measures a unitary construct: Cronbach's alpha was .97, and the item-scale total correlations ranged from .67 to .83. Of the 25 items on the WZ-PTSD scale, 20 also appear on the BSI.

Signal Detection Analyses

To evaluate the relative diagnostic utility of the different PTSD measures, signal detection analyses were conducted for each scale, using the methods detailed by Kraemer (1992). The WZ-PTSD scale, the CR-PTSD scale, the Mississippi Scale, and the PK scale were the primary measures analyzed. In addition, we analyzed the shortened version of the WZ-PTSD scale that appears in the BSI. Finally, to compare the performance of the WZ-PTSD with general measures of distress we analyzed the Global Severity Index (GSI) of the SCL-90-R, the *F* scale of the MMPI, and a 25-item scale of randomly selected SCL-90-R items. By chance, this latter scale of 25 random SCL-90-R items shared 6 items with the WZ-PTSD scale.

Signal detection analyses were conducted on the total derivation sample of 202 subjects, with 198 subjects available for analyses involving the PK Scale. However, since only 166 subjects had completed either the full MMPI or the full MMPI-2, analyses involving the *F* scale were based on this reduced sample size.

Results

Figure 1 displays ROC curves for the WZ-PTSD scale, the Mississippi Scale, and the PK scale, allowing a comparison of their performance across all possible cutoff scores. The ROC's for all three questionnaires are lo-

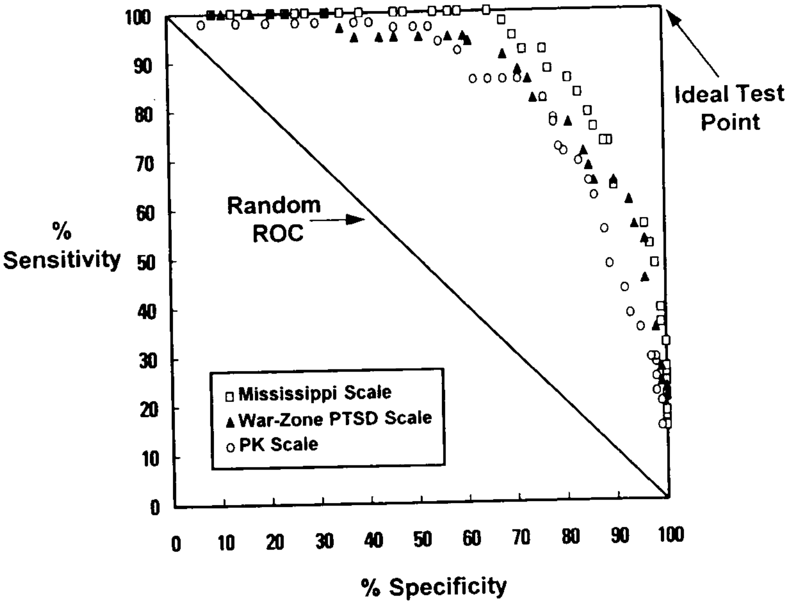

Figure 1. Receiver Operating Characteristic (ROC) curves for three measures of war-zone-related PTSD.

cated well above the random ROC. This shows that all three measures are correlated with the PTSD diagnosis. The ROC for the Mississippi Scale extends the farthest toward the ideal test point in the upper right corner, indicating that across a range of cutoff scores it is the best predictor of PTSD. The Spearman rank point-biserial correlation (r_{pb}) between the Mississippi Scale and the PTSD diagnosis, which is proportional to the area under the ROC curve and serves as a measure of the overall quality of a scale (see Kraemer, 1988), was .69. The r_{pb} was .62 for the WZ-PTSD scale and .57 for the PK scale.

Figure 2 displays the QROC curves for the same questionnaires. These curves, derived by plotting the quality of sensitivity against the quality of specificity, are a one-to-one remapping of the ROC curves (Kraemer, 1992). The main advantage of this remapping is that the optimally sensitive and specific cutoffs for each scale can be identified visually and compared to each other. The optimally sensitive cutoff for a scale is the highest point on the graph (or in case of a tie, the highest point that is farthest to the

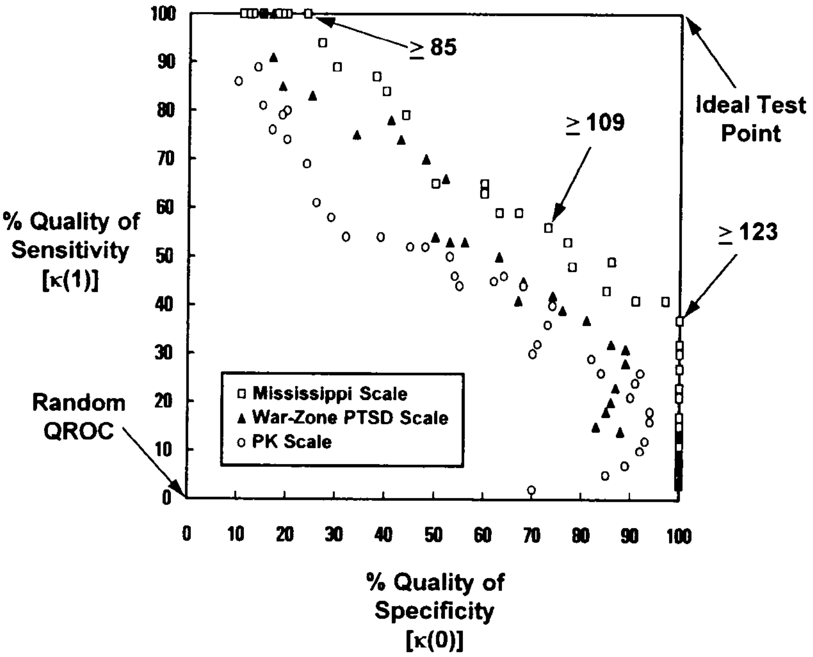

Figure 2. Quality Receiver Operating Characteristic (QROC) curves for three measures of war-zone-related PTSD.

right), and the optimally specific cutoff for a scale is the point farthest to the right (or in case of a tie, the point farthest to the right that is the highest). Identifying optimally efficient cutoffs from QROC curves is more complicated, but in general they will lie on or near the main diagonal running from the random QROC at the zero point to the ideal test point (Kraemer, 1992).

Inspection of Figure 2 also shows that the Mississippi Scale was superior to the WZ-PTSD scale and the PK scale in terms of calibrated sensitivity, specificity, and efficiency. The optimally sensitive cutoff on the Mississippi Scale was 85, which had perfect sensitivity and a 24% quality of specificity [κ(0)]. The optimally specific cutoff was 123, which had perfect specificity and a 37% quality of sensitivity [κ(1)]. The optimally efficient cutoff was 109, which is very close to the cutoff of 107 suggested by Keane et al. (1988) for use in clinical populations. Interestingly, cutoff scores in the range of 101 to 109 had virtually identical quality of efficiency, indicating that the scores in this range provide comparable diagnostic utility.

Table 1 presents, for each scale in the study: (a) the r_{pb} between the scale and the diagnosis, representing the overall quality of the scale; (b) the optimally efficient cutoff score; (c) the level of the test (the proportion of test positives); (d) its sensitivity, specificity, and efficiency; and (e) the kappa coefficient representing the quality of efficiency [κ(.5)].

The scales in Table 1 are listed in the order of their quality of efficiency [κ(.5)]. In order to determine if the observed differences among the scales were statistically significant, we conducted pairwise comparisons using a jackknife procedure (Bloch & Kraemer, 1989; Efron, 1982) for testing the difference between two correlated kappa coefficients (i.e., kappas ob-

Table 1. The Diagnostic Utility of Several Measures of War-Zone-Related PTSD and Global Distress Based on Derivation Sample (N = 202, Base Rate = 67%)[a]

Scale	r_{pb} with Diagnosis	Cutoff	Level of Test	Sensitivity	Specificity	Efficiency	κ(.5)
Mississippi Scale	.69	109	.61	.83	.83	.83	.63
WZ-PTSD Scale	.62	1.3	.72	.90	.65	.82	.58
WZ-PTSD Scale/BSI	.61	1.3	.71	.89	.65	.81	.56
CR-PTSD Scale	.57	1.3	.67	.85	.70	.80	.55
PK Scale	.57	26	.60	.78	.78	.78	.54
Global Severity Index	.56	1.1	.69	.86	.67	.80	.53
25 random SCL-90-R items	.54	1.1	.68	.85	.65	.78	.50
F Scale	.40	67	.71	.82	.57	.75	.40

[a]*Note:* r_{pb}= Spearman rank point-biserial correlation representing overall quality; diagnosis = interview-based diagnosis of PTSD; cutoff = optimally efficient cutoff score; level test = proporton of test positives; κ(.5) = kappa coefficient representing quality of efficiency.

tained from the same cases). Because we predicted greater efficiency for the PTSD scales relative to global measures of distress, but did not make *a priori* assumptions about relative efficiency among the PTSD scales, we used one-tailed tests to compare PTSD scales with measures of global distress, and two-tailed tests to compare two PTSD scales. We did not test differences among the measures of global distress.

With respect to the quality of efficiency, the jackknife analyses revealed that the Mississippi Scale significantly exceeded the F scale ($p <$ .005), the PK scale ($p < .01$), and the 25-item random scale ($p < .05$). The WZ-PTSD scale exceeded the F scale and the 25-item random scale (p's $< .05$), whereas the CR-PTSD scale, the PK scale, and the shortened version of the WZ-PTSD scale exceeded only the F scale (p's $< .05$).

Table 2 presents the performance and quality of the optimally sensitive and specific cutoff scores for the Mississippi Scale, the PK scale, the full and shortened versions of the WZ-PTSD scale, and the CR-PTSD scale. Complete information on these additional cutoffs for the other scales in the study can be obtained by contacting the first author. These data indicate that, by choosing very low cutoffs on these scales (thus increasing the level of the test), perfect or nearly perfect sensitivity can be obtained, although at a cost of a large number of false positives. Similarly, by choosing very high cutoffs (thus decreasing the level of the test), perfect or nearly perfect specificity can be obtained, at a cost of a large number of false negatives.

Table 2. Optimally Sensitive and Specific Cutoff Scores for the WZ-PTSD Scale, Mississippi Scale, PK Scale, and CR-PTSD Scale Based on Derivation Sample (N = 202, Base Rate = 67%)[a]

Scale	Cutoff	Level of Test	Sensitivity	Specificity	$\kappa(1)$	$\kappa(0)$
		Optimally Sensitive Cutoffs				
Mississippi Scale	85	.90	1.00	.32	1.00	.24
WZ-PTSD Scale	.40	.93	1.00	.23	1.00	.17
WZ-PTSD Scale/BSI	.30	.95	1.00	.17	1.00	.12
CR-PTSD Scale	.30	.93	.99	.21	.90	.15
PK Scale	8	.93	.99	.20	.89	.14
		Optimally Specific Cutoffs				
Mississippi Scale	123	.44	.65	1.00	.37	1.00
WZ-PTSD Scale	3.1	.21	.32	1.00	.13	1.00
WZ-PTSD Scale/BSI	3.3	.13	.19	1.00	.07	1.00
CR-PTSD Scale	3.0	.13	.19	1.00	.07	1.00
PK Scale	37	.28	.41	.98	.18	.94

[a] *Note*: cutoff = optimally sensitive or specific cutoff score; level of test = proportion of test positives; $\kappa(1)$ = kappa coefficient representing quality of sensitivity; $\kappa(0)$ = kappa coefficient representing quality of specificity.

Study 2

Method

Subjects

In order to cross-validate the initial findings, we evaluated the diagnostic utility of the scales used in the derivation phase in a new sample of 99 Vietnam theater veterans. Subjects contacted the National Center for PTSD from November, 1991 to September, 1992 either to obtain clinical services or to participate in research. The percentage of subjects seeking clinical services was slightly higher in the cross-validation sample (65% vs. 63%) and the base rate of PTSD was slightly lower (61% vs. 67%). As in the derivation phase, 75% of the PTSD diagnoses were made using the SCID PTSD module and 25% were made using the CAPS.

The cross-validation sample was quite similar demographically to the derivation sample with a few exceptions: The cross-validation sample included fewer Whites and more African-Americans ($p < .05$) as well as fewer veterans of the Marines and more veterans of the Air Force and Navy ($p < .05$). No other significant differences were found between the two samples.

Signal Detection Analyses

All of the signal detection analyses conducted in the derivation phase were repeated with the cross-validation sample. As expected, the optimal cutoffs in the cross-validation sample differed from the optimal cutoffs in the derivation sample for many of the scales evaluated. A prediction equation is optimal for the sample from which it was derived, and is usually less than optimal in new samples. The purpose of cross-validation is to determine how much predictive power is retained when the prediction equation is used in a new sample. Accordingly, in the cross-validation sample we examined the performance and quality only of those cutoffs determined to be optimal for the derivation sample.

Results

Table 3 describes the performance and quality in the cross-validation sample of the optimally efficient cutoffs identified in Study 1. The table presents for each scale considered: (a) the r_{pb} between the scale and the diagnosis, representing the overall quality of the scale; (b) the optimally

Table 3. The Diagnostic Utility of Several Measures of War-Zone-Related PTSD and Global Distress Based on Cross-Validation Sample (N = 99, Base Rate = 61%)[a]

Scale	r_{pb} with Diagnosis	Cutoff	Level of Test	Sensitivity	Specificity	Efficiency	$\kappa(.5)$
Mississippi Scale	.68	109	.52	.77	.82	.79	.58
WZ-PTSD Scale	.62	1.3	.64	.87	.72	.81	.59
WZ-PTSD Scale/BSI	.61	1.3	.64	.85	.69	.79	.55
CR-PTSD Scale	.59	1.3	.58	.76	.69	.73	.45
PK Scale	.54	26	.49	.68	.81	.73	.47
Global Severity Index	.58	1.1	.64	.83	.67	.77	.51
25 random SCL-90-R items	.55	1.1	.61	.79	.69	.76	.48
F Scale	.49	67	.64	.81	.62	.73	.44

[a] *Note:* r_{pb} = Spearman rank point-biserial correlation coefficient representing overall quality; diagnosis = interview-based diagnosis of PTSD; cutoff= optimally efficient cutoff derived in Study 1; level of test = proportion of test positives; $\kappa(.5)$ = kappa coefficient representing quality of efficiency.

efficient cutoff score identified in Study 1; (c) the level of the test; (d) the sensitivity, specificity, and efficiency; and (e) $\kappa(.5)$.

As indicated by the r_{pb}'s, all of the scales retained nearly the same overall quality in the cross-validation sample as they demonstrated the derivation sample, although most optimal cutoffs identified in Study 1 had lower quality of efficiency in Study 2. The exception was the WZ-PTSD scale, which was stable across the derivation and cross-validation samples, even showing a slight increase in quality in the cross-validation sample. Furthermore, the same cutoff on the WZ-PTSD scale (1.3) was the optimally efficient cutoff in both Study 1 and Study 2, which was not the case with any other scale. The optimally efficient cutoffs identified in Study 2 for the remaining scales can be obtained from the first author.

Jackknife analyses comparing the scales on quality of efficiency revealed that the Mississippi Scale exceeded the CR-PTSD scale and the PK scale (p's < .05), and the WZ-PTSD scale exceeded the CR-PTSD scale (p < .01). No other comparisons were significant, in part because of the reduced sample size.

Table 4 presents the performance and quality in the cross-validation sample of the optimally sensitive and specific cutoff scores identified in Study 1. In general, the optimally sensitive cutoffs from the Study 1 continued to show excellent sensitivity as well as somewhat higher specificity. Also, with the exception of the Mississippi Scale, which showed a decline in specificity, the optimally specific cutoffs from Study 1 continued to show excellent specificity, although with somewhat lower sensitivity.

Table 4. Optimally Sensitive and Specific Cutoff Scores for the WZ-PTSD Scale, Mississippi Scale, PK Scale, and CR-PTSD Scale Based on Cross-Validation Sample (N = 99, Base Rate = 61%)[a]

Scale	Cutoff	Level of Test	Sensitivity	Specificity	κ(1)	κ(0)
		Optimally Sensitive Cutoffs				
Mississippi Scale	85	.80	.98	.45	.91	.31
WZ-PTSD Scale	.40	.88	.98	.28	.86	.18
WZ-PTSD Scale/BSI	.30	.89	1.00	.28	1.00	.19
CR-PTSD Scale	.30	.89	1.00	.28	1.00	.19
PK Scale	8	.87	1.00	.35	1.00	.25
		Optimally Specific Cutoffs				
Mississippi Scale	123	.33	.52	.93	.28	.77
WZ-PTSD Scale	3.1	.13	.22	1.00	.10	1.00
WZ-PTSD Scale/BSI	3.3	.12	.20	1.00	.09	1.00
CR-PTSD Scale	3.0	.11	.19	1.00	.08	1.00
PK Scale	37	.18	.28	1.00	.13	1.00

[a] *Note:* cutoff = optimally sensitive or specific cutoff score derived in Study 1; level of test = proportion of test positives; κ(1) = kappa coefficient representing quality of sensitivity; κ(0) = kappa coefficient representing quality of specificity.

Discussion

In this investigation we developed the WZ-PTSD scale, an empirically derived measure of war-zone-related PTSD. We then used signal detection analyses to evaluate the relative diagnostic utility of the WZ-PTSD scale, other measures of PTSD, and measures of global distress. Finally, we cross-validated the results in a new sample of Vietnam veterans.

Results from Study 1 indicated the potential value of the WZ-PTSD scale as a measure of war-zone-related PTSD. First, although PTSD subjects scored significantly higher than non-PTSD subjects on almost all SCL-90-R items, the 25 items retained for the WZ-PTSD scale displayed particularly robust discrimination between PTSD and non-PTSD subjects in two different subsamples. These 25 items were strongly intercorrelated and the internal consistency of the WZ-PTSD scale was quite high, suggesting that it measures a unitary construct.

Interestingly, only 11 of the 25 items chosen for the WZ-PTSD scale also appear on the CR-PTSD scale. This low level of overlap, which motivated our decision to develop a separate scale for war-zone-related PTSD, may be due to a number of factors including the differential impact of criminal victimization versus combat, gender differences in response to trauma, other unidentified demographic differences in the samples on which the two scales were derived, or measurement error.

Second, when the diagnostic utility of the various scales in the study was evaluated on the full derivation sample, the WZ-PTSD scale had the second-highest quality of efficiency, exceeded only by the Mississippi Scale. The WZ-PTSD scale demonstrated significantly greater quality of efficiency relative to the 25-item random scale and to the *F* scale. These results suggest that the WZ-PTSD scale is related specifically to the construct of PTSD and has incremental predictive value over these measures of global distress.

Study 2 provided further evidence of the value of the WZ-PTSD scale for the diagnosis of PTSD. When the optimally efficient cutoff scores identified in Study 1 were applied to a new sample of veterans, all scales except the WZ-PTSD scale showed some reduction in utility. The WZ-PTSD scale showed a slight increase in utility, displaying the highest cross-validated quality of efficiency of any of the scales investigated. In this new sample, the quality of efficiency of the WZ-PTSD scale significantly exceeded that of the CR-PTSD scale, suggesting that the WZ-PTSD scale has greater diagnostic utility in a population of combat veterans. Finally, unlike other scales, the optimally efficient cutoff score for the WZ-PTSD scale was the same in both the derivation and cross-validation samples.

However, it is important to note that the WZ-PTSD scale did not differ significantly from the GSI in the quality of efficiency in either the derivation sample or the cross-validation sample. This may be due in part to the high levels of distress among the non-PTSD subjects. Approximately 55% of the non-PTSD subjects in the two studies met criteria for a current Axis I disorder other than PTSD, and virtually all of them met criteria for a lifetime Axis I disorder. In clinical samples such as ours, it may be that no subset of SCL-90-R items could outperform the GSI, a measure of global distress which is likely to be affected by high levels of comorbid diagnoses and functional impairment. In a less distressed population, such as a community sample of combat veterans whose symptoms are limited primarily to PTSD, the WZ-PTSD might have greater incremental utility. Further research in different war-zone-exposed populations is needed to test this possibility. Nonetheless, the WZ-PTSD scale had a higher quality of efficiency in both the derivation and the cross-validation studies, and this replication is evidence that the difference between these two scales is genuine and stable. From a practical assessment perspective, even a slight improvement in classification accuracy can be meaningful.

Taken together, the results indicate that the diagnostic utility of the WZ-PTSD scale is stable and generalizes beyond the derivation sample. Although the generalizability of the performance and quality of the cutoffs will be limited to the extent that new populations and settings are similar to the ones examined in this study, this still constitutes a significant demonstration of psychometric integrity.

Apart from the development of the WZ-PTSD scale, this study makes two other contributions to the literature on the assessment of PTSD. First, this is one of the few studies evaluating the relative diagnostic utility of multiple measures of PTSD in the same sample of subjects. In the largest such study, Kulka et al. (1990) compared the utility of the Mississippi Scale, the PK scale, and the IES, finding the Mississippi Scale to be the best predictor of a PTSD diagnosis. The present study replicates and extends the NVVRS findings, evaluating not only multiple PTSD scales but also several measures of global distress. Such head-to-head comparisons are essential for identifying optimal measures for assessing PTSD and replications in different settings with different types of trauma survivors would be valuable to establish the generalizability of the present findings.

Second, this is the first study to use signal detection methods to investigate the diagnostic utility of PTSD questionnaires. The signal detection approach is valuable both for identifying optimally sensitive, specific, and efficient cutoff scores on a given questionnaire, as well as for comparing the diagnostic utility of different questionnaires. The data from this study can guide the selection of PTSD scales and cutoffs for different clinical and research applications.

In summary, the results are promising in terms of the usefulness of the WZ-PTSD Scale, scored from either the SCL-90-R or the BSI. It is internally consistent and its diagnostic utility meets or exceeds that of other established PTSD scales. Because the WZ-PTSD scale is derived from such a widely used parent instrument, its availability will allow the assessment of PTSD in a variety of clinical and research contexts.

Acknowledgments

We wish to thank Danny Kaloupek for his helpful comments on earlier versions of this manuscript. Portions of this study were presented at the 1992 meeting of the International Society for Traumatic Stress Studies, Los Angeles, CA. Jennifer Huska is now at the Department of Psychology, Indiana University.

References

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed. rev.). Washington, DC: Author.

Blake, D. D., Keane, T. M., Wine, P. R., Mora, C., Taylor, K. L., & Lyons, J. A. (1990). Prevalence of PTSD symptoms in combat veterans seeking medical treatment. *Journal of Traumatic Stress*, *3*, 15-27.

Blake, D. D., Weathers, F. W., Nagy, L. N., Kaloupek, D. G., Klauminzer, G., Charney, D. S., & Keane, T. M. (1990). A clinician rating scale for assessing current and lifetime PTSD: The CAPS-1. *The Behavior Therapist*, *13*, 187-188.

Bloch, D. A., & Kraemer, H. C. 2 x 2 kappa coefficients: Measures of agreement or association. *Biometrics*, *45*, 269-287.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory (MMPI-2): Manual for administration and scoring*. Minneapolis: University of Minnesota Press.

Davidson, J., & Smith, R. (1990). Traumatic experiences in psychiatric outpatients. *Journal of Traumatic Stress*, *3*, 459-475.

Derogatis, L. R. (1983). *SCL-90-R: Administration, scoring, & procedures manual-II*. Towson, MD: Clinical Psychometric Research.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Foa, E. B., Riggs, D. S., Dancu, C. V., & Rothbaum, B.O. (1993). Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. *Journal of Traumatic Stress*, *6*, 459-473.

Hammarberg, M. (1992). Penn Inventory for posttraumatic stress disorder: Psychometric properties. *Psychological Assessment*, *4*, 67-76.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29-36.

Hathaway, S. R., & McKinley, J. C. (1983). *The Minnesota Multiphasic Personality Inventory manual*. New York: Psychological Corporation.

Hays, W. L. (1988). *Statistics*. NY: Holt, Rinehart, and Winston.

Herman, D. S., Weathers, F. W., Litz, B. T., Joaquim, S. G., & Keane, T. M. (1993, October). *The PK scale of the MMPI-2: Reliability and validity of the embedded and stand-alone versions*. Paper presented at the annual meeting of the International Society for Traumatic Stress Studies, San Antonio.

Horowitz, M. J., Wilner, N., & Alvarez, W. (1979). Impact of Event Scale: A measure of subjective stress. *Psychosomatic Medicine*, *41*, 209-218.

Keane, T. M., Caddell, J. M., & Taylor, K. L. (1988). Mississippi Scale for combat-related posttraumatic stress disorder: Three studies in reliability and validity. *Journal of Consulting and Clinical Psychology*, *56*, 85-90.

Keane, T. M., Fairbank, J. A., Caddell, J. M., Zimering, R. T., Taylor, K. L., & Mora, C. A. (1989). Clinical evaluation of a measure to assess combat exposure. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *1*, 53-55.

Keane, T. M., Kolb, L. C., & Thomas, R. T. (1990). *A psychophysiological study of chronic PTSD*. Department of Veterans Affairs Cooperative Study #334.

Keane, T. M., Malloy, P. F., & Fairbank, J. A. (1984). Empirical development of an MMPI subscale for the assessment of combat-related posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*, *52*, 888-891.

Keane, T. M., Wolfe, J., & Taylor, K. L. (1987). Post-traumatic stress disorder: Evidence for diagnostic validity and methods of psychological assessment. *Journal of Clinical Psychology*, *43*, 32-43.

Kraemer, H. C. (1988). Assessment of 2 x 2 associations: Generalization of signal-detection methodology. *The American Statistician*, *42*, 37-49.

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines*. Newbury Park, CA: Sage Publications.

Kulka, R. A., Schlenger, W. E., Fairbank, J. A., Hough, R. L., Jordan, B. K., Marmar, C. R., & Weiss, D. S. (1990). *Trauma and the Vietnam war generation: Report of findings from the National Vietnam Veterans Readjustment Study*. New York: Brunner-Mazel.

Litz, B. T., Penk, W. E., Walsh, S., Hyer, L., Blake, D. D., Marx, B., Keane, T. M., & Bitman, D. (1991). Similarities and differences between MMPI and MMPI-2 applications to the assessment of posttraumatic stress disorder. *Journal of Personality Assessment*, *57*, 238-253.

Lyons, J. A., & Keane, T. M. (1992). Keane PTSD Scale: MMPI and MMPI-2 update. *Journal of Traumatic Stress*, *5*, 111-117.

McFall, M. E., Smith, D. E., Roszell, D. K., Tarver, D. J., & Malas, K. L. (1990). Convergent validity of measures of PTSD in Vietnam combat veterans. *American Journal of Psychiatry*, *147*, 645-648.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: Its history, characteristics, and validity. *Archives of General Psychiatry*, *38*, 318-389.

Saunders, B. E., Arata, C. M., & Kilpatrick, D. G. (1990). Development of a crime-related posttraumatic stress disorder scale for women within the Symptom Checklist-90-Revised. *Journal of Traumatic Stress*, *3*, 439-448.

Spiro, A., Schnurr, P. P., & Aldwin, C. M. (1994). Combat-related PTSD symptoms in older men. *Psychology and Aging*, *9*, 17-26.

Spitzer, R. L., Williams, J. B. W., Gibbons, M., & First, M. B. (1990). *Structured Clinical Interview for DSM-III-R*. Washington, DC: American Psychiatric Press.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.

Watson, C. G. (1990). Psychometric posttraumatic stress disorder measurement techniques: A review. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*, 460-469.

Weathers, F. W., Blake, D. D., Krinsley, K. E., Haddad, W. H., Huska, J. A., & Keane, T. M. (1992, November). *The Clinician-Administered PTSD Scale: Reliability and construct validity*. Paper presented at the annual meeting of the Association for Advancement of Behavior Therapy, Boston, MA.

Weathers, F. W., & Litz, B. T. (1994). Psychometric properties of the Clinician-Administered PTSD Scale - Form 1 (CAPS-1). *PTSD Research Quarterly*, *5*, 2-6.

Weathers, F. W., Litz, B. T., Herman, D. S., Huska, J. A., & Keane, T. M. (1993, October). *The PTSD Checklist (PCL): Reliability, validity, and diagnostic utility*. Paper presented at the annual meeting of the International Society for Traumatic Stress Studies, San Antonio.

Zilberg, N. J., Weiss, D. S., & Horowitz, M. J. (1982). Impact of Event Scale: A cross-validation study and some empirical evidence supporting a conceptual model of stress response syndromes. *Journal of Consulting and Clinical Psychology*, *50*, 407-414.